

Step-by-step key to Day 1 exercises.

Exercise #1: Retrieve the sequence of the BAC LB5-28F9 from GenBank

Address <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=nucleotide>

NCBI

Entrez PubMed Nucleotide Protein Genome

Search Nucleotide for LB5-28F9 Go Clear

Limits Preview/Index History Clipboard

Display Summary Show: 20 Send to Text

1: [AC144989](#)
Callicebus moloch clone LB5-28F9, WORKING DRAFT SEQUENCE
gi|32441307|gb|AC144989.2|[32441307]



1: [AC144989](#). Reports Callicebus moloch...[gi:32441307]

LOCUS AC144989 221500 bp DNA linear HTG 03-JUL-2003
DEFINITION Callicebus moloch clone LB5-28F9, WORKING DRAFT SEQUENCE.
ACCESSION AC144989
VERSION AC144989.2 GI:32441307
KEYWORDS HTG; HTGS_PHASE2; HTGS_DRAFT.
SOURCE Callicebus moloch (Dusky titi)
ORGANISM [Callicebus moloch](#)
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; [Primates](#); Platyrrhini; Cebidae; Callicebinae;
Callicebus.

Species information

Exercise #2: Retrieve a subsequence 160000-221500 and reverse complement it

Display default Show: 20 Send to File **Get Subsequence** Features

1: [AC144989](#). Callicebus moloch...[gi:32441307]

LOCUS AC144989 221500 bp DNA linear HTG 03-JUL-2003
DEFINITION Callicebus moloch clone LB5-28F9, WORKING DRAFT SEQUENCE.
ACCESSION AC144989
VERSION AC144989.1
KEYWORDS HTG
SOURCE Callicebus moloch
ORGANISM Callicebus moloch
REFERENCE 1
AUTHORS Cheng, Z., Kozlov, I., and Rubin, E.H.
TITLE Direct Submission

Change Sequence Range

From 160000 to 221500 (1-221500)
☒ Reverse Complement

Set Range Unset Range Dismiss

Done Internet

Exercise #2: Convert the extracted sequence to FASTA format

Entrez PubMed Nucleotide Protein Genome

Search Nucleotide for Go Clear

Display FASTA Show: 20 Send to File Get Subsequence Features

1: [AC144989](#). Callicebus moloch...[gi:32441307]

LOCUS AC144989 61501 bp DNA linear HTG 03-JUL-2003
DEFINITION Callicebus moloch clone LB5-28F9, WORKING DRAFT SEQUENCE.
ACCESSION [AC144989](#) REGION: complement(160000..221500)
VERSION AC144989.2 GI:32441307
KEYWORDS HTG: HTGS PHASE2: HTGS DRAFT.

And save as a text document on your computer.

Display FASTA Show: 20 Send to File Get Subsequence

1: [AC144989](#). Callicebus moloch...[gi:32441307]

Save As

Save in: Desktop

My Recent Documents Desktop My Documents My Computer My Network Places

File name: MYNAME.TXT

Save as type: Document

Save Cancel

>gnl|pgaberk|
CGCGGAATTCAG
CCTAAGCAGGAGC
GCTTTATTGCAAC
CTTAGTCGGCAC
AGCAGGGGCTGG
TGTTTAGTGTGCC
GGGACAGAGTCC
TTCTGTGCAAAA
GTGAAGGCTCC
TAAACATCACAA
ATGAGCCCAATT
AATATTAGTAAA
CTGGTGAATAAG
GGCAGGCAAAAT
AAAAATACAAAA
AAAAATCGCTTG
AACAAGAGCAAA
ACCCCCATCCCG
GCAGGCTCTGCA
CCAGTGCTAAAT
ACAAAAGGTAAA
AACTGTGCAAA
AAACACTGGCAA
GGCGGATCACAA
ACAAAAAATTAGCTTGGCATGGTGTTCATTCCTATAATCCAGCTACTTGGGAGACTGAGACAGGAGAA

Exercise #3: Find what coding sequences are present in your extracted sequence

Use Megablast

Address  <http://www.ncbi.nlm.nih.gov/BLAST/>

 **BLAST**

PubMed Entrez **BLAST** OMIM Taxonomy Structure

Info

- FAQs
- News
- References
- NCBI
- Contribut

Education

- Program selection guide
- Tutorial
- URL API guide

Download

- Databases
- Documentation
- Executables
- Source code

Support

- Helpdesk
- Mailing list

NEW 12 May 2004 BLAST 2.2.9 has been released. [Read more...](#)

Nucleotide

- Discontiguous megablast
- **Megablast**
- Nucleotide-nucleotide BLAST (blastn)
- Search for short, nearly exact matches
- Search trace archives with megablast or discontiguous megablast

Protein

- Protein-protein BLAST (blastp)
- PHI- and PSI-BLAST
- Search for short, nearly exact matches
- Search the conserved domain database (rpsblast)
- Search by domain architecture (cdart)

Translated

- Translated query vs. protein database (blastx)
- Protein query vs. translated database (tblastn)
- Translated query vs. translated database (tblastx)

Genomes

- Chicken, cow, pig, dog, sheep, cat **NEW**
- Environmental samples
- Human, mouse, rat
- Fugu rubripes, zebrafish
- Insects, nematodes, plants, fungi, malaria
- Microbial genomes, other eukaryotic genomes

Special

- Search for gene expression data (GEO BLAST)
- Align two sequences (bl2seq)
- Screen for vector contamination (VecScreen)
- Immunoglobulin BLAST (IgBlast)

Meta

- Retrieve results by RID
- Get this page with javascript-free links



Exercise #3: How to run Megablast

[Search](#)

Load query file from disk:

[Set subsequence](#) From: To:

[Choose database](#)

Return alignment endpoints only ☐

Now: or

LOAD input sequence here (FASTA format)

Options for advanced blasting

since we know that our sequence is most closely related to human, we limit the search to human entries only

[Limit by entrez query](#) or select from:

[Choose filter](#) ☒ Low complexity ☒ Human repeats ☐ Mask for lookup table only ☐ Mask lower case

[Expect](#)

[Word Size](#)

a long word size makes searching more efficient for long sequences

Exercise #4: BLAST Results for BAC sequence

Query: your BAC sequence - Subject: database hit



results of BLAST

Subject accession #: look in the NUCLEOTIDE database for information on these sequences (sse next page)

e-value: significance of the alignment. Alignments with the lowest e-value are the most significant. However, remember that long alignments have lower e-values

```
#BLASTN 2.2.9 [May-01-2004]
# Query: Callicebus moloch clone L
# Database: nr
# Fields: Query id, Subject id, % identity, alignment length, mismatches, gap openings, q. start, q. end, s. start, s. end, e-value, bit score
```

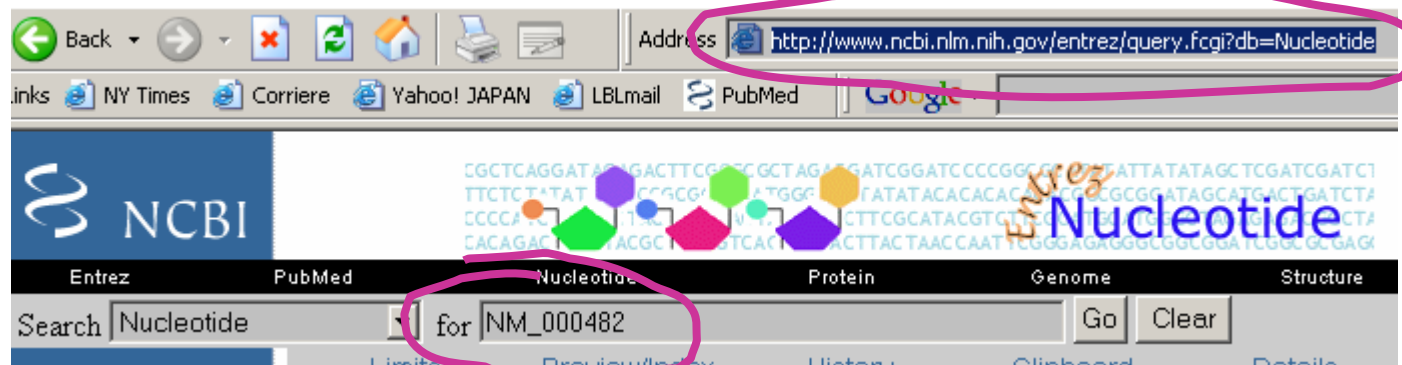
Callicebus	gi 9625348 gb AC074203.3	89.58	4817	320	110	47821	52528	35822	40565	0.0	6191
Callicebus	gi 9625348 gb AC074203.3	90.65	1797	117	29	52562	54324	40588	42379	0.0	2475
Callicebus	gi 9625348 gb AC074203.3	88.95	1883	136	37	43350	45192	31651	33501	0.0	2352
Callicebus	gi 9625348 gb AC074203.3	88.74	1235	84	33	46617	47823	34590	35797	0.0	1519

Callicebus	gi 178756 gb J02758.1 HUMAPOA4A	89.22	306	20	6	20149	20450	3307	3607	3e-105	396
Callicebus	gi 22091457 ref NM_052968.2	93.28	982	60	5	45271	46247	371	1351	0.0	1502
Callicebus	gi 22091457 ref NM_052968.2	90.66	503	29	10	46617	47117	1381	1867	0.0	679
Callicebus	gi 22091457 ref NM_052968.2	96.55	116	4	0	44451	44566	55	170	3e-46	200
Callicebus	gi 22091457 ref NM_052968.2	95.00	120	6	0	45073	45192	173	292	5e-45	196
Callicebus	gi 6707432 gb AF202889.1	93.28	982	60	5	45271	46247	371	1351	0.0	1502

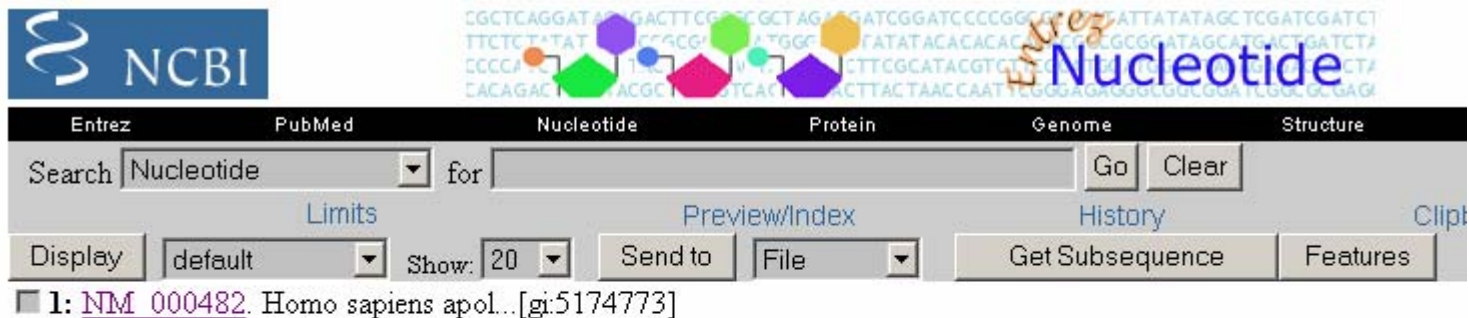
3 RefSeq hits: this is a coding sequence (mRNA)

Callicebus	gi 49902035 gb BC074764.1	93.26	89	6	0	17677	17765	1282	1194	4e-27	137
Callicebus	gi 5174773 ref NM_000482.2	91.94	707	53	4	19254	19958	522	1226	0.0	1027
Callicebus	gi 5174773 ref NM_000482.2	92.73	165	9	3	17603	17765	1	164	2e-59	244
Callicebus	gi 5174773 ref NM_000482.2	96.88	128	4	0	18124	18251	163	290	4e-53	223
Callicebus	gi 563319 gb M10373.1 HUMAPOAIVA	91.95	708	51	6	19254	19958	6	710	0.0	1027
Callicebus	gi 178778 gb M14566.1 HUMAPOAIV	91.94	707	53	4	19254	19958	348	1052	0.0	1027
Callicebus	gi 178778 gb M14566.1 HUMAPOAIV	86.98	192	11	4	20149	20337	1108	1292	4e-53	223
Callicebus	gi 178778 gb M14566.1 HUMAPOAIV	97.41	116	3	0	18136	18251	1	116	6e-48	206
Callicebus	gi 178758 gb M13654.1 HUMAPOA4B	91.95	708	51	6	19254	19958	459	1163	0.0	1027
Callicebus	gi 178758 gb M13654.1 HUMAPOA4B	96.88	128	4	0	18124	18251	100	227	4e-53	223
Callicebus	gi 178758 gb M13654.1 HUMAPOA4B	95.05	101	5	0	17665	17765	1	101	9e-36	166
Callicebus	gi 28771 emb X02162.1 HSAPOAIB	90.83	665	53	8	5427	6087	287	947	0.0	919
Callicebus	gi 28771 emb X02162.1 HSAPOAIB	91.03	156	5	2	4688	4842	129	283	2e-55	231
Callicebus	gi 4557320 ref NM_000039.1	90.80	663	53	8	5427	6085	239	897	0.0	915
Callicebus	gi 4557320 ref NM_000039.1	91.03	156	5	2	4688	4842	81	235	2e-55	231
Callicebus	gi 128722 gb M22825.1 HUMAPOAIB	90.80	663	53	8	5427	6085	220	878	0.0	915

Exercise #4: Find out what are the genes present in your BAC sequence by looking in the “nucleotide” database



search the nucleotide database for the accession # found in your BLAST search



LOCUS NM_000482 1466 bp mRNA linear PRI 20-DEC-2003

DEFINITION Homo sapiens apolipoprotein A-IV (APOA4), mRNA.

ACCESSION NM_000482

VERSION NM_000482.2 GI:5174773

KEYWORDS .

SOURCE Homo sapiens (human)

ORGANISM Homo sapiens

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.

Exercise #5: Use BL2seq to locate the intron-exon boundaries of your predicted coding sequences

(TIP: work with multiple browser windows to copy and paste information)

Address  <http://www.ncbi.nlm.nih.gov/BLAST/>



BLAST

PubMed	Entrez	BLAST	OMIM	Taxonomy	Structure
Info NEW 12 May 2004 BLAST 2.2.9 has been released. Read more...					
Info <ul style="list-style-type: none">• FAQs• News• References• NCBI Contributors Education <ul style="list-style-type: none">• Program selection guide• Tutorial• URL API guide Download <ul style="list-style-type: none">• Databases• Documentation• Executables• Source code Support <ul style="list-style-type: none">• Help• Feedback	Nucleotide <ul style="list-style-type: none">• Discontiguous megablast• Megablast• Nucleotide-nucleotide BLAST (blastn)• Search for short, nearly exact matches• Search trace archives with megablast or discontiguous megablast		Protein <ul style="list-style-type: none">• Protein-protein BLAST (blastp)• PHI- and PSI-BLAST• Search for short, nearly exact matches• Search the conserved domain database (rpsblast)• Search by domain architecture (cdart)		
	Translated <ul style="list-style-type: none">• Translated query vs. protein database (blastx)• Protein query vs. translated database (tblastn)• Translated query vs. translated database (tblastx)		Genomes <ul style="list-style-type: none">• Chicken, cow, pig, dog, sheep, cat NEW• Environmental samples• Human, mouse, rat• Fugu rubripes, zebrafish• Insects, nematodes, plants, fungi, malaria• Microbial genomes, other eukaryotic genomes		
	Special <ul style="list-style-type: none">• Search for gene expression data (GEO BLAST)• Align two sequences (bl2seq)• Screen for vector contamination (VecScreen)• Immunoglobulin BLAST (IgBlast)		Meta <ul style="list-style-type: none">• Retrieve results by RID• Get this page with javascript-free links		



Exercise #5: Compare your BAC DNA sequence with the sequence of the human APOA5 cDNA

blastn is the program to compare DNA to DNA

The screenshot shows the NCBI BLASTN web interface. The 'Program' dropdown is set to 'blastn' and is circled in pink. The 'Matrix' dropdown is set to 'Not Applicable'. The 'Parameters used in BLASTN program only:' section shows 'Reward for a match' set to 1 and 'Penalty for a mismatch' set to -1. A pink arrow points from the text 'modify these parameters from the default settings to obtain sequence alignments surrounding intron-exon junctions' to the mismatch penalty field. The 'Strand option' is set to 'Both strands'. The 'Open gap' is set to 3 and 'extension gap' is set to 1, with a pink arrow pointing from the text 'and extension gap penalties' to the extension gap field. The 'gap x_dropoff' is set to 50, 'expect' is 10.0, and 'word size' is 11. The 'Filter' checkbox is checked. The 'Align' button is visible. The 'Sequence 1' section has a text input field containing 'PO_BACfor Exercise1.txt' and a 'Browse...' button, both circled in pink. The 'Sequence 2' section has a text input field containing 'NM_052968' and a 'Browse...' button, with the input field circled in pink. At the bottom, there are 'Align' and 'Clear Input' buttons.

Program: **blastn** Matrix: Not Applicable

Parameters used in **BLASTN** program only:

Reward for a match: 1 Penalty for a mismatch: **-1** modify these parameters from the default settings to obtain sequence alignments surrounding intron-exon junctions

☐ Use **Mega BLAST** Strand option: Both strands

Open gap: **3** and extension gap: **1** penalties

gap x_dropoff: 50 expect: 10.0 word size: 11 Filter: ☒ Align

Sequence 1 Enter accession or GI: **PO_BACfor Exercise1.txt** or download from file: **PO_BACfor Exercise1.txt** Browse...

or sequence in FASTA format from: 0 to: 0

one of the RefSeq accession numbers from your Megablast

Sequence 2 Enter accession or GI: **NM_052968** or download from file: Browse...

or sequence in FASTA format from: 0 to: 0

Align Clear Input

Exercise #5: Use BL2seq to locate the intron-exon boundaries of your predicted coding sequences

Intron-Exon junction rule:

exon1

GT.....AG

exon2

BLAST 2 SEQUENCES RESULTS VERSION BLASTN 2.2.10

Sequence 1 gnl|pgaberk|T022-28F9:c221500-160000 Callicebus moloch clone LB5-28F9, WORKING DRAFT
SEQUENCE Length 61501 (1 .. 61501)

Sequence 2 gi 22091457 Homo sapiens apolipoprotein A-V (APOA5), mRNA Length 1889 (1 .. 1889)

Intron sequences are in bold and red. Look for consensus splice sites: GT-intron-AG (GT and AG are part of the intron)

Exon 3

Query: 45065 **cccag**gagcctgaaagacagccttgagcaagacctcaacaatatgaacaagttcctggaa 45124 ← your BAC sequence
||| 
Sbjct: 165 cccgagaccctgaaagacagccttgagcaagacctcaacaatatgaacaagttcctggaa 224 ← human APOA5 cDNA
apolipoprotein AV 50 P A T L K D S L E Q D L N N M N K F I E ← human APOA5 protein sequence
.....
Query: 46201 cactcaccaggctttgcaaaccagctttccagtgctcatttgggaattctcataa**gttg** 46260
||| 
Sbjct: 1305 cattcaccaggctttgcaaaccagctccagtgctcatttgggaatgctcatgagtta 1364

Exon 4

Query: 46601 **a**tgctcctttcaag-----gggagtagggagggagaaaggcaccatgcatgtgggtga 46655
||| 
Sbjct: 1360 agttactccattcaagggtgagggagtagggagggag--aggcaccatgcatgtgggtga 1417
.....
Query: 47075 gaagcctagacttctggctcaaataaattagatgtttatgatagaa 47120
|||
Sbjct: 1825 gaagcctagacttctggctcaaataaattagatgtttatgataaaa 1870

Exon 2

Query: 44451 **cag**cggttttcggccaccagggcacggaaaggcttctgggactacttccgccagaccagcg 44510
|||
Sbjct: 55 cagcggttttcggccaccagggcacggaaaggcttctgggactacttccgccagaccagcg 114
apolipoprotein AV 13 S A F S A T Q A R K G F W D Y F S Q T S
.....
Query: 44511 gggacaaaggcaggatggagcagatccatcagcagaagatggctcgtgaacccg**g** 44566
||| 
Sbjct: 115 gggacaaaggcagggtggagcagatccatcagcagaagatggctcgcgagcccgcg 170
apolipoprotein AV 33 G D K G R V E Q I H Q Q K M A R E P A

Exon 1

Query: 44265 agcagacaatggcaagcatggctgtcgtgctcacctgggctctggctctcctctcag 44321
|||
Sbjct: 1 agcaggtaatggcaagcatggctgcccgtgctcacctgggctctggctcttcttctcag 57
apolipoprotein AV 1 M A A V L T W A L A L L S

Exercise #6: Use EXTRACTSEQ to extract the sequence of your predicted C. moloch APOA5 cDNA

bioweb.pasteur.fr/seqanal/interfaces/extractseq.html

EXTRACTSEQ : Extract regions from a sequence (EMBOSS)

your e-mail

(● = required, ● = conditionally required)

Input section

● sequence -- any [single sequence] (-sequence) : please enter either :

1. the name of a **file**:

2. or the **actual data** here:

(sequence format)

[Return to the main part with your favorite browser's Back function]

Required section

enter the exon containing intervals here

● Regions to extract (eg: 4-57,78-94) (-regions)

[Return to the main part with your favorite browser's Back function]

the file "outseq.out"
contains the extracted
sequence (text document)

Exercise #6: UseGETORF on your EXTRACTSEQ output to verify that the cDNA you extracted contains an intact open reading frame

EXTRACTSEQ : Extract regions from a sequence (EMBOSS)

Results:

[extractseq.out](#)

[outseq.out](#) (1.98 Kb)

[standard error file](#)

From now, this files will remain accessible for 10 days at: <http://bioweb.pasteur.fr/seqanal/tmp/extractseq/A49843010995272/>

You can save them individually by the **Save file** function if needed.

Exercise #7: Retrieve the human APOA5 protein sequence . There are several ways to do this:

1. You can find that information in the GenBank (nucleotide database) page for the mRNA accession #: look for RefSeq product, this is the protein sequence accession number

CDS

```
/w_... 18..1109
/ gene="APOA5"
/ note="regeneration-associated protein 3; apolipoprotein
A5;
go_function: lipid binding [goid 0008289] [evidence IEA];
go_process: lipid transport [goid 0006869] [evidence IEA]"
/codon_start=1
/product="apolipoprotein A-V"
/protein_id="NP_443200.1"
/db_xref="GI:16445025"
/db_xref="GeneID:116519"
/db_xref="LocusID:116519"
/db_xref="MIM:606368"
/translation="MAAVLTWALALLSAFSATQARKGFWDYFSQTS GDKGRVEQIHQQ
...
```

2. You can query the ENTREZ gene database with the mRNA accession # (<http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi>). Look for results in the "Gene" section.
Open the gene page and for the NP_XXXXXX accession number

Entrez Gene

Med Nucleotide Protein Genome Structure

for [] Go Clear

Limits Preview/Index History Clipboard Details

Display Graphics Show: 5 Send to Text

1: APOA5 apolipoprotein A-V [*Homo sapiens*]
GeneID: 116519 Locus tag: [HGNC:17288](#); [MIM: 606368](#)
Transcripts and products: (shown on reverse complement genome) [RefSeq below](#)

NC_000011

116200236 5' 116197738 3'

[NM_052968](#) [NP_443200](#)

■ - coding region ■ - untranslated region

Exercise #7: compare the APOA5 cDNA sequence of your BAC and of human protein:

blastx is the program to compare a DNA sequence to a protein sequence

The image shows the NCBI BLAST web interface. A pink oval highlights the 'Program' dropdown menu, which is set to 'blastx'. An arrow points from the text 'blastx is the program to compare a DNA sequence to a protein sequence' to this dropdown. Another pink oval highlights the 'Matrix' dropdown menu, which is set to 'BLOSUM62'. Below these, the text 'Parameters used in BLASTN program only:' is visible. The 'Reward for a match:' and 'Penalty for a mismatch:' fields are empty. A checkbox for 'Use Mega BLAST' is unchecked, and the 'Strand option' is set to 'Not Applicable'. The 'Open gap' is set to 11 and the 'extension gap' is set to 1. The 'gap x_dropoff' is set to 50, 'expect' is 10.0, and 'word size' is 3. The 'Filter' checkbox is checked. The 'Align' button is highlighted with a pink oval. Below the 'Align' button, the 'Sequence 1' section has a text input field containing 'your cDNA', which is also highlighted with a pink oval. The 'Sequence 2' section has a text input field containing 'human protein ACC#', which is also highlighted with a pink oval. At the bottom, there are 'Align' and 'Clear Input' buttons.

Program **blastx** Matrix **BLOSUM62**

Parameters used in [BLASTN](#) program only:

Reward for a match: Penalty for a mismatch:

☐ Use [Mega BLAST](#) Strand option **Not Applicable**

Open gap and extension gap penalties

gap x_dropoff [expect](#) word size [Filter](#) ☒ **Align**

Sequence 1 Enter accession or GI or download from file **your cDNA**

or sequence in FASTA format from: to:


Sequence 2 Enter accession or GI or download from file **human protein ACC#**

or sequence in FASTA format from: to:

Align **Clear Input**

Exercise #7: Blastx alignment of Callicebus and human APOA5

Score = 676 bits (1745), Expect = 0.0
Identities = 340/363 (93%), Positives = 352/363 (96%)
Frame = +3



Query: 18 MAVVLTWALALLSAFSATQARKGFWDYFRQTS GDKGRMEQIHQQKMAREPASLKDSLEQD 197
MA VLTWALALLSAFSATQARKGFWDYF QTS GDKGR+EQIHQQKMAREPA+LKDSLEQD
Sbjct: 1 MAAVLTWALALLSAFSATQARKGFWDYFSQTS GDKGRVEQIHQQKMAREPATLKDSLEQD 60

Query: 198 LNMNMKFLERLGPLSGSEAPRIPREPVGMRQQLQEELEEVRARLQPHMAEAHEL VGWNLE 377
LNMNMKFLE+L PLSGSEAPR+P++PVGMR+QLQEELEEVR+ARLQP+MAEAHEL VGWNLE
Sbjct: 61 LNMNMKFLEKLRLPLSGSEAPRLPQDPVGMRRQLQEELEEVRKARLQPYMAEAHEL VGWNLE 120

Query: 378 GLRQQLKPYTMDLMEQVALRVQELQEQLRVVGEDTKAQLLG VGEARALLQELQSRVVHH 557
GLRQQLKPYTMDLMEQVALRVQELQEQLRVVGEDTKAQLLG V EA ALLQ LQSRVVHH
Sbjct: 121 GLRQQLKPYTMDLMEQVALRVQELQEQLRVVGEDTKAQLLG VDEAWALLQGLQSRVVHH 180

Query: 558 TGRFKELFHPYAESLVSGIGRHHVQELHRSVAPHAPASPARLSRCVQVLSRKLT LKAKALH 737
TGRFKELFHPYAESLVSGIGRHHVQELHRSVAPHAPASPARLSRCVQVLSRKLT LKAKALH
Sbjct: 181 TGRFKELFHPYAESLVSGIGRHHVQELHRSVAPHAPASPARLSRCVQVLSRKLT LKAKALH 240

Query: 738 ARIQQMLDQLREELSRAFA GTGAEQAGPDPQMLSEEVQRQLQA FRQDTYLQIAAFTRAI 917
ARIQQMLDQLREELSRAFA GTG E+GAGPDPQMLSEEVQRQLQA FRQDTYLQIAAFTRAI
Sbjct: 241 ARIQQMLDQLREELSRAFA GTGTEEGAGPDPQMLSEEVQRQLQA FRQDTYLQIAAFTRAI 300

Query: 918 DQETEEVQQQLAPPPPGHSAFAPEFGQMSDKALSKLQARLDDLWEDITYSLHDQGHSHL 1097
DQETEEVQQQLAPPPPGHSAFAPEF Q DS K LSKLQARLDDLWEDIT+SLHDQGHSHL
Sbjct: 301 DQETEEVQQQLAPPPPGHSAFAPEFQQTDSGKVL SKLQARLDDLWEDITHSLHDQGHSHL 360

Query: 1098 GEP 1106
G+P
Sbjct: 361 GDP 363

aminoacid change
with a chemically different aminoacid

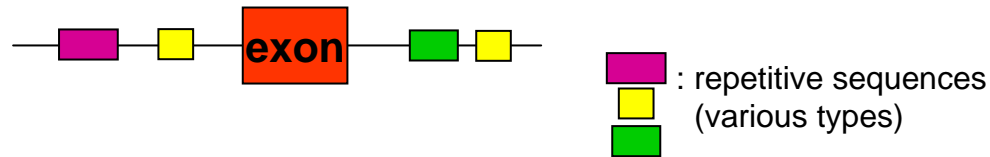
conserved
aminoacid

aminoacid change
with a chemically similar aminoacid
(for example, negatively charged
with negatively charged)

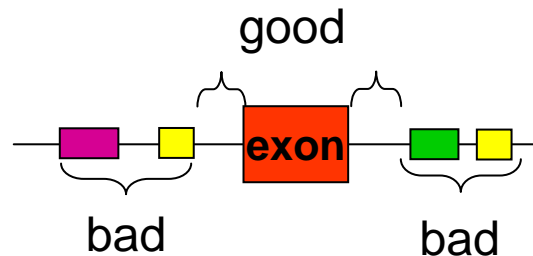
Exercise #8-11: design PCR primer for resequencing exons

Primers to amplify exons for resequencing of clinical samples

- Primers need to be designed in the intronic sequences surrounding the target exon.
- The goal is to pick primers that will only amplify the target region and not other parts of the genome.
- This is largely achieved by avoiding picking primers in repetitive regions of the genome.
- Because of poor sequencing quality near priming sites, it is good practice to design primers at least 30-50 bp away from the exon.

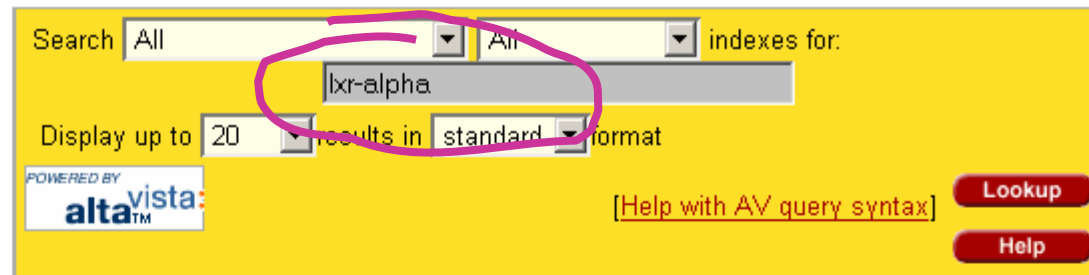


primer design:



Exercise #8: retrieve the sequence of LXR-alpha exon 4 and of its 2 surrounding introns.

Use the ENSEMBL genome browser: www.ensembl.org/



1 matches in the ***Homo sapiens*** Gene index [first 5 matches shown]:

1. Ensembl Gene: [ENSG00000025434](http://www.ensembl.org:80/Homo_sapiens/geneview?gene=ENSG00000025434)

Ensembl gene ENSG00000025434 has 2 transcripts: ENST00000344715, ENST00000298843

Oxysterols receptor **LXR-alpha** (Liver X receptor alpha) (Nuclear orphan receptor **LXR-alpha**). [Source:Uniprot/SWISSPROT;Acc:Q13133]

The gene has the following external identifiers mapped to it:

GO: GO:0003707, GO:0004887, GO:0005634, GO:0003713, GO:0006355, GO:0003700

HUGO: NR1H3, 7966

LocusLink: 10062

MIM: 602423

protein_id: AAH08819.1, AAA85856.1

RefSeq: NP_005684, NM_005693

Uniprot/SWISSPROT: NRH3_HUMAN, Q13133

 http://www.ensembl.org:80/Homo_sapiens/geneview?gene=ENSG00000025434

1 matches in the ***Mus musculus*** Gene index [first 5 matches shown]:

1. Ensembl Gene: [ENSMUSG00000002108](http://www.ensembl.org:80/Mus_musculus/geneview?gene=ENSMUSG00000002108)

Ensembl gene ENSMUSG00000002108 has 1 transcript: ENSMUST00000002177

Oxysterols receptor **LXR-alpha** (Liver X receptor alpha) (Nuclear orphan receptor **LXR-alpha**). [Source:Uniprot/SWISSPROT;Acc:Q9Z0Y9]

The gene has the following external identifiers mapped to it:

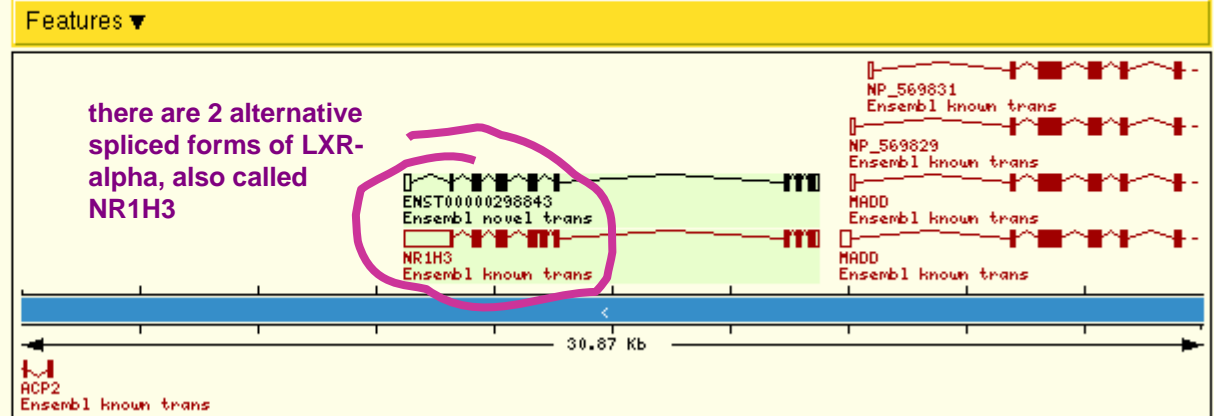
EMBL: AJ132601, AJ132600, AJ132599, AF085745

LocusLink: 22259

Exercise #8: retrieve the sequence of LXR-alpha exon 4 and of its 2 surrounding introns.

Transcript Structure

- 1: ENST00000298843 [\[Transcript information\]](#) [\[Exon information\]](#) [\[Protein information\]](#)
 2: [NR1H3](#) (ENST00000344715) [\[Transcript information\]](#) [\[Exon information\]](#) [\[Protein information\]](#)



Intron 7-8	11	1	47246154	47246387	234 bp	gtgtggaggaggggcaatgggaaac.....cctcaactctcttggt
8 ENSE00000713314	11	1	47246388	47246482	95 bp	ACCGGCCCAACGTGCAGGACCAGCTCCAGGTAGAGAGGCTGCAGCACACA CCCTGCATGCCTACGTCTCCATCCACCATCCCCAT
Intron 8-9	11	1	47246483	47246676	194 bp	gtgagttctccccatggtgttctctt.....gtgtttgtctctctcc
9 ENSE00001215906	11	1	47246677	47246965	289 bp	GACCGACTGATGTTCCACGGATGCTAATGAAACTGGTGAGCCTCCGGACC GTCCACTCAGAGCAAGTGTTCGACTGCGTCTGCAGGACAAAAAGCTCCCC TCTGAGATCTGGGATGTGCACGAATGACTGTTCTGTCCCCATATTTCTGT CGGATGGCTGAGGCCTGGTGGCTGCCTCCTAGAAGTGGAAACAGACTGAGA TTCCTGGGAGCTGGGCAAGGAGATCCTCCCGTGGCATTAAAAGAGAGTC aaaggggttgcgagttttgtggctactgagcagtgaggccctcgctaacac.
3' downstream sequence						

KEY: -Up/downstream region
 -Intron sequence
 -UTR region

Display bases either side of intron.

Display bases of 5' 3' flanking sequence.

[View full intron sequence](#)

[View only exon sequence](#)

Redraw

Select
 "Redraw"


Exercise #8: retrieve the sequence of LXR-alpha exon 3 and of its 2 surrounding introns.

Length Sequence

<u>Intron 2-3</u>		11	1	47237387	47237917	531 bp	gtaagcttcattccatccctctccctgagcccagaccgcaggctccacgcctcctgtag gaatcagcctccttcattacctgccttcttccctccagagagcagtcagagtcatt cttagtggtgcttgccctcccgcccagatcacctctccctgggtccagtgccctggccct gcaggcaccgcggcagtcctccctcagctctggatttgctgctagagagtggtccagctgag tgcttacctgctctggctttgaagagttttatctgatctctgaaatgcatacactccag ccccccaaagggacaaggattaacatcttcatttaaggctcctgagatgtaagaaactaca agtgactagtcctagctagagcccacacagactctagggtcccaaagcctgagctgggac tttgctgccctctaagggtggggataagttgcagtttcccagctaggacgctggggcgt ggagccgggatggggcctgagaccccttgctgcctctctctttggagctcag
3	<u>ENSE00000839256</u>	11	1	47237918	47238106	189 bp	ACTCTGCGGTGGAGCTGTGGAAGCCAGGCCACAGGATGCAAGCAGCCAGGCCAGGGAG GCAGCAGCTGCATCCTCAGAGAGGAAGCCAGGATGCCCCACTCTGCTGGGGGTACTGCAG GGGTGGGGCTGGAGGCTGCAGAGCCCACAGCCCTGCTCACCAGGGCAGAGCCCCCTTCAG AACCACAG
<u>Intron 3-4</u>		11	1	47238107	47238535	429 bp	gtgaggagcttctgggtttggaggaggtaggggtccagattccaggtcctggatctggaa gaggttccttgggggtttttactttatatataatctcatggttaagttcagaggcttag agctaactaaatctgactgatctaagtgtgaattttgtctctaggccttctgagcctca cttctctgtttataaaatggaataaaaaattatggttgctcataaggatcagtgcatata aaagggtcatacagtagctagaacataatggcacttggcaaatgagggctactcttctca taaaagagagactggagtttgtataatgaaggggaatgaaggctcactgagtgccaggga gtggctgagtcagggaacatgatgttttctcgggggagagcgttgaagcactttcc tgtatccag

copy these sequences to 1 text file and save on your desktop

Exercise #9: use RepeatMasker www.repeatmasker.org/cgi-bin/WEBRepeatMasker to identify repeats in your sequence



RepeatMasker Web Server

[RepeatMasker](#) screens DNA sequences in fasta format against a library of repetitive elements and reports the results as a table annotating the masked regions.
Reference: A.F.A. Smit & P. Green, unpublished data. Current Version: 3.0.2

[Check Current Queue Status](#)

Basic Options

[Large sequences](#) will be queued, and may take a while to process.

Enter the [file](#) to process:

Or paste the sequence(s) in [FASTA-format](#):

```
gtgaggagctctctgggtttggaggaggtaggggtccagattccagggtcctggatctggaa
cgggttccttgggggttttactttatatataaatctcatggttaagttcagaggcttttag
agctaactaaatctgactgactctaagtgtgaattttgtctctaggcctttctgagcctca
ctttccttgtttataaaaatggaaaataaaaattatggttgtcataaggatcagtgcatata
aaaggctcatcacagtacctaagaacataaatggcacttggcaaatgagggctactcttctca
taaaagagagactggagtttgtataatgaaggggaatgaagggtcactgagtggtcaggga
gtggctgagtcaggggagaacatgatgttttctcctcgggggagagcgttgaagcactttcc
```

Select [return format](#): ☒ html ☐ tar file ☐ links

Select return method: ☒ html ☐ email

Advanced Options

[Speed/Sensitivity](#): ☐ rush ☐ quick ☒ default ☐ slow

[DNA source](#):

your sequence

output options

Exercise #9: RepeatMasker results for human LXR-alpha intron 2/3, exon 3 and intron 3/4

Repeat Annotations:

SW score	perc div.	perc del.	perc ins.	query sequence	position in query begin end (left)	matching repeat repeat class/family	position in repeat begin end (left)	ID
377	29.1	15.7	0.0	UnnamedSequence	157 290 (971)	+ MIRb SINE/MIR	106 260 (8)	1
2300	9.6	0.7	0.0	UnnamedSequence	824 1115 (146)	+ AluSx SINE/Alu	1 294 (18)	2

Masked Sequence:

>UnnamedSequence

```

GTGAGGAGCTTCTGGGTTTGGAGGAGGTAGGGGTCCAGATTCCAGGTCCT
GGATCTGGAAGAGGTTTCCTTGGGGGTTTTTACTTTATATATAATCTCATG
GTAAAGTTCAGAGGCTTTAGAGCTAACTAAATCTGACTGATCTAAGTGTG
AATTTTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
CTCTTCTCA
TAAAAGAGAGACTGGAGTTTGTATAATGAAGGGAATGAAGGTCAGTGA
GCCCAGGGCAGTGGCTGAGTCAGGGAGAACATGATGTTTTTCTCGGGGG
AGAGCGTTGAAGCACTTTCTGTATCCAGATCCGTCCACAAAAGCGGA
AAAAGGGGCCAGCCCCAAAATGCTGGGGAACGAGCTATGCAGCGTGTGT
GGGGACAAGGCCTCGGGCTTCCACTACAATGTTCTGAGCTGCGAGGGCTG
CAAGGGATTCTTCCGCCGACGCTCATCAAGGGAGCGCACTACATCTGCC
ACAGTGCGCGGCCACTGCCCCATGGACACCTACATGCGTCGCAAGTGCCAG
GAGTGCGGCTTCGCAAAATGCCGTCAGGCTGGCATGCGGGAGGAGTGTGA
GTTTCTGGGGCTGGAGTGGGGAAGAGGCTGAGGGGAAAGAGGGGGCCAGG
GTGTGACCCAAAACAGGTGCCTGAACCTTGCAGGGGCTAACTGATCCCTAA
GTATGGATCCCAGTATCTTTCTNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
CTTCTTGCCCTTTACCCAGTGCTGTCTGCTTTTCT
GGAGCCCCAAACCACCCCTTTGCCCATCCTTCCCTCCTGTCTTTCCCC
CACCCCTTGCCCCATCCTTTCCCCATCTGCTCCCTTCCCTCATATTTGGC
CCTGTCCTTAG
  
```

position of the repeat within the sequence

exon sequence

masked repeat

repeat type

Exercise #10: Use Primer3 frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi
to design primers to amplify your exon-containing sequence

Primer3

pick primers from a DNA sequence

[disclaimer](#)

[cautions](#)

Paste source sequence below (5'→3', string of ACGTNacgtm -- other letters treated as N -- numbers and blanks ignored). FASTA sequence (vector, ALUs, LINEs, etc.) or use a [Mispriming Library \(repeat library\)](#):

```
GCCCAGGGCAGTGGCTGAGTCAGGGAGAAACATGATGTTTTCTCGGGGG
AGAGCGTTGAAGCACTTTCTGTATCCAG[AGATCCGTCCACAAAAGCGGA
AAAAGGGGCCAGCCCCCAAAATGCTGGGGAACGAGCTATGCAGCGTGTGT
GGGGACAAGGCCTCGGGCTTCCACTACAATGTTCTGAGCTGCGAGGGCTG
CAAGGGATTCTTCCGCCGAGCGTCATCAAGGGAGCGCACTACATCTGCC
ACAGTGGCGGCCACTGCCCATGGACACCTACATGCGTCGCAAGTGCCAG
GAGTG]TCGGCTTCGCAAAATGCCGTCAGGCTGGCATGCGGGAGGAGTGTGA
```

**you can mark the boundary of the exon
junction in the sequence with []
for exon resequencing, leave 50 extra
bases on both sides**

☒ Pick left primer or use left
primer below.

☐ Pick hybridization probe
(internal oligo) or use oligo below.

☒ Pick right primer or use right
primer below (5'→3' on opposite
strand).

General Primer Picking Conditions

[Primer Size](#) Min: Opt: Max:

[Primer Tm](#) Min: Opt: Max: [Max Tm Difference](#):

[Product Tm](#) Min: Opt: Max:

[Primer GC%](#) Min: Opt: Max:

[Max Self Complementarity](#): [Max 3' Self Complementarity](#):

[Max #N's](#): [Max Poly-X](#):

[Inside Target Penalty](#): [Outside Target Penalty](#): [Set Inside Target Penalty to](#)

[First Base Index](#): [CG Clamp](#):

[Salt Concentration](#): [Annealing Oligo Concentration](#): [\(Not the concentration of c](#)

☒ [Liberal Base](#) ☐ [Show Debugging Info](#) ☒ Do not treat ambiguity codes in libraries as consensus

Pick Primers

Reset Form

**default parameters are
normally good**

**Exercise #11: Use BLAST for short, nearly exact matches (see BLAST page)
to verify that your primers are unique in the human genome**

[Search](#)

```
CAGTGGCTGAGTCAGGGAGA  
CTTTCCCCTCAGCCTCTTC
```

you can BLAST forward and reverse primers at once

[Set subsequence](#)

From: To:

[Choose database](#)

Now:

BLAST! or

Options for advanced blasting

[Limit by entrez
query](#)

or select from:

[Choose filter](#)

☐ Low complexity ☐ Human repeats ☐ Mask for lookup table only ☐ Mask lower case

[Expect](#)

[Word Size](#)

these parameters are optimized for short sequence matches:


- short sequences have higher expect values even for perfect matches
- word size 7 (default value is normally 11) facilitates identifying short sequence matches

[Other advanced](#)

Exercise #11: Inspect BLAST output: you want to see that your primers only hit your target region

Sequences producing significant alignments:		Score (bits)	E Value
gi 21622769 gb AC090589.9 	Homo sapiens chromosome 11, clon...	<u>40</u>	0.071
gi 24850147 gb AC018410.24 	Homo sapiens chromosome 11, clo...	<u>40</u>	0.071
gi 29124047 gb AC010427.5 	Homo sapiens chromosome 5 clone ...	<u>38</u>	0.28
gi 28827851 gb AC026740.6 	Homo sapiens chromosome 5 clone ...	<u>38</u>	0.28
gi 21955075 gb AC105935.2 	Homo sapiens chromosome 3 clone ...	<u>36</u>	1.1

correct chromosome

■ >[gi|21622769|gb|AC090589.9|](#)  Homo sapiens chromosome 11, clone RP11-390K5, complete sequence
Length = 190017


Score = 40.1 bits (20), Expect = 0.071
Identities = 20/20 (100%)
Strand = Plus / Plus

Score = 38.2 bits (19), Expect = 0.28
Identities = 19/19 (100%)
Strand = Plus / Minus

Query: 1 cagtggctgagtcagggaga 20
 |||||
Sbjct: 150797 cagtggctgagtcagggaga 150816

Query: 21 ctttccccctcagcctcttc 39
 |||||
Sbjct: 151177 ctttccccctcagcctcttc 151159

wrong chromosome

■ >[gi|29124047|gb|AC010427.5|](#)  Homo sapiens chromosome 5 clone CTD-2198K18, complete sequence
Length = 68883

Score = 38.2 bits (19), Expect = 0.28
Identities = 19/19 (100%)
Strand = Plus / Minus

Query: 16 ggagactttccccctcagcc 34
 |||||
Sbjct: 8525 ggagactttccccctcagcc 8507

this match is a mix of the forward and reverse primer and can be ignored